Autonomous Vocal and Backing Track Mixing

Kelian (Mike) Li

Introduction

Automatic mixing has become an emerging field in recent years. With many studies attend to create a general-purposed solution, this work presents an approach to perform automatic vocal mixing using convolutional neural networks. Proposed new mixing parameter extraction methods and source separation expand the amount of training data for automatic level balance, compression, equalization, and reverberation. A baseline model is implemented combining data analysis and rule-based approaches. The listening test shows that the baseline model is comparable with the deep learning model. Analysis of the subjective testing result provides insights on the perceptual preferences of music mixing. The processing order is:

Equalization => Compression => Reverberation => Level Balance

Data Collection

the vocals and the backing tracks are extracted from 21936 clips of 30-second audio in a subset of the Million Song Dataset. This data is used in level balance and compression. For equalization and reverberation, the dataset is the 100 full-length songs from the train set of MUSDB18HQ.

Level Balance*:* Relative integrated loudness between the vocal and the backing track is chosen as the loudness measure.

Compression*:* EBU loudness range of the vocal signal is the dynamic range measure in this work. An iterative process optimizes the vocal loudness range by modifying the compression threshold and compression ratio. The compression.

Equalization*:* The purely rule-based baseline equalization does not require any data. For the deep learning training data, the main idea is to create ”raw” vocals by applying randomized EQ on the processed vocals. 4 out of 9 center frequencies in [63, 125, 250, 500, 1K, 2K, 4K, 8K, 16K] Hz are randomly chosen. Random gain changes between [-15, 15] dB are applied on the processed vocals.

Reverberation*:* Blind reverb estimation remains an unsolved problem in public research. In this work, a commercial plugin Chameleon is utilized to directly extract the estimated reverb impulse responses from the wet vocal signal. The next step is to match the reverb parameters by the genetic algorithm

Baseline Method

Level Balance: The baseline system sets the relative integrated loudness of -1.77dB between the vocals and the backing tracks which is the average from the 21936 clips in the separated Million Song Dataset.

Compression: The average loudness range of the processed vocals in the separated Million Song Dataset is 16.36dB which is set as the targeted dynamic range for iterative process described above.

Equalization: The baseline automatic equalization is achieved by a rule-based frequency masking reduction method.

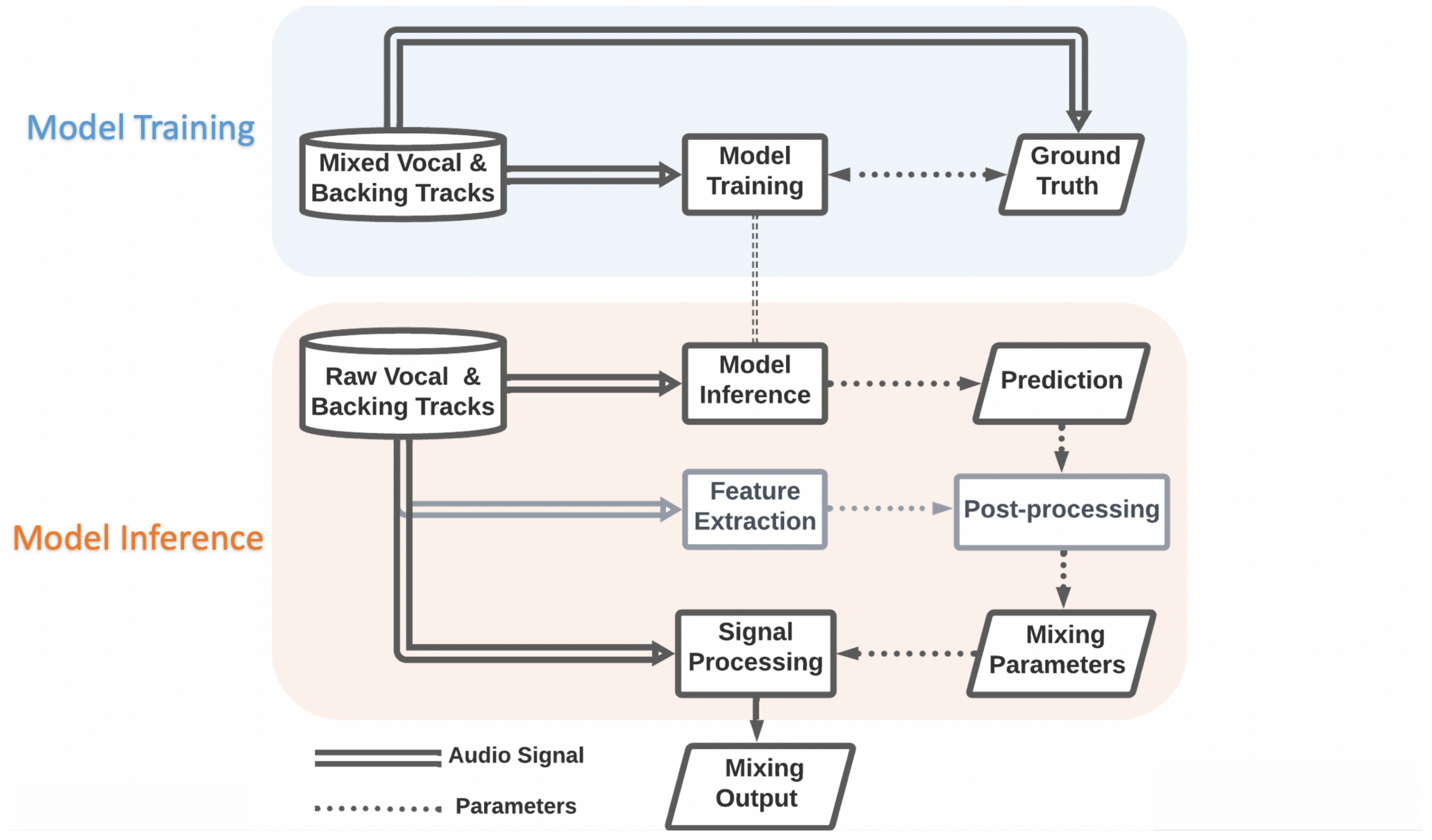
Reverberation: The baseline automatic reverberation combines the data analysis and rule-based approaches. Our rule-based reverberation linearly maps the tempo to the reverb time. Other reverb parameters set to mean values extracted from the train set of MUSDB18HQ.

CNN Model

Four independent convolutional neural network regression models are trained to predict the mixing parameters from the mel-spectrogram of the vocal and the backing track. As mentioned in the data collection, the peak filters are applied on 4 out 9 center frequencies which leads to unbalanced training data. To tackle that, the loss function has to be altered and prevents the model from outputting zeros only. A modified loss function is

$$L_{EQ} = l_{changed} + 0.1 \cdot l_{unchanged}$$

where l*changed* is the L2 loss of the 4 frequency bands where the ground truth values are changed, and l*unchanged* is the loss of the 5 frequency bands where the ground truth values are zeros. During the inference, the top 4 frequency bands which have the largest absolute gain change prediction are selected.

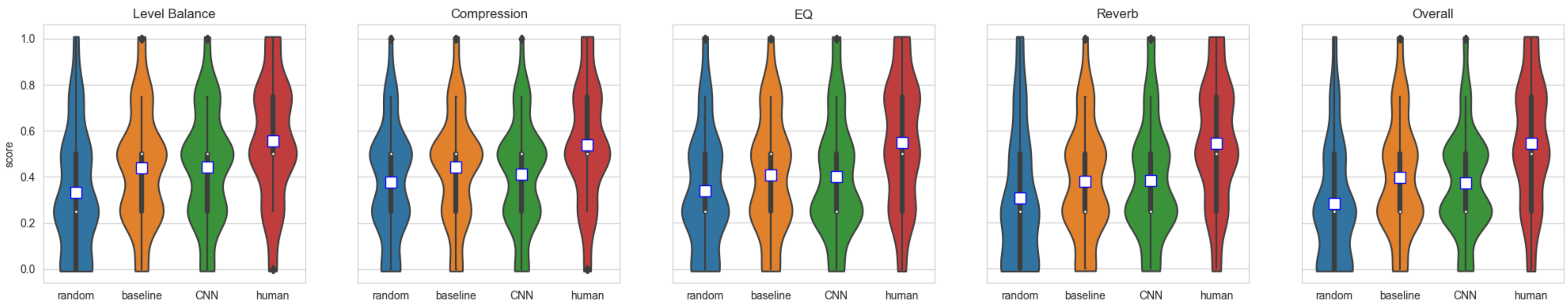


For objective evaluation. the CNN model can outperform the baseline system in level balance and compression. The CNN model can only perform slightly better than the baseline model on most of the reverb parameters. The EQ model successfully converges during the training, but the validation result shows that it lacks the ability of generalization.

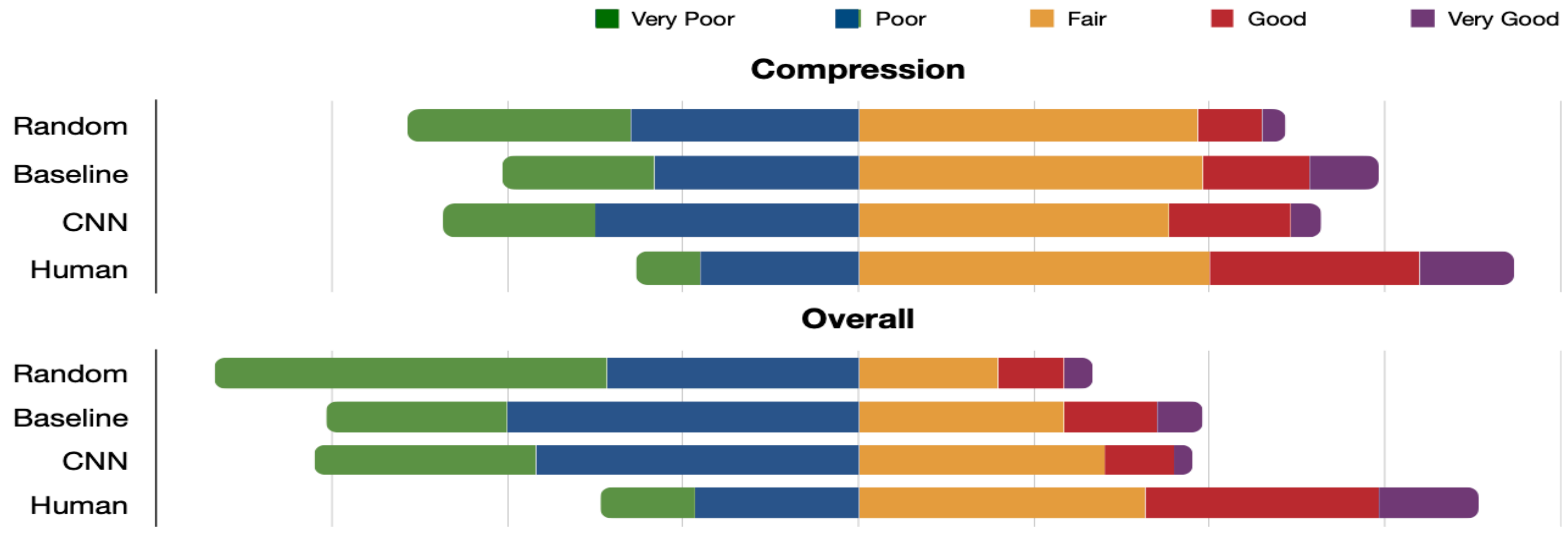
	relative loudness	loudness range	EQ gain	dry/wet ratio	reverb time	FT	RS	LCF	LCG	LCQ	HCF	HCG	HCQ
CNN	**1.64**	**2.63**	4.48	**6.13**	**1.006**	**0.401**	**7.30**	41.55	**3.62**	**0.355**	**3390**	**5.01**	**0.14**
mean	2.13	2.88	**3.33**	7.16	1.007	0.403	7.31	**40.26**	3.65	0.360	3462	5.35	0.17

Listening Test

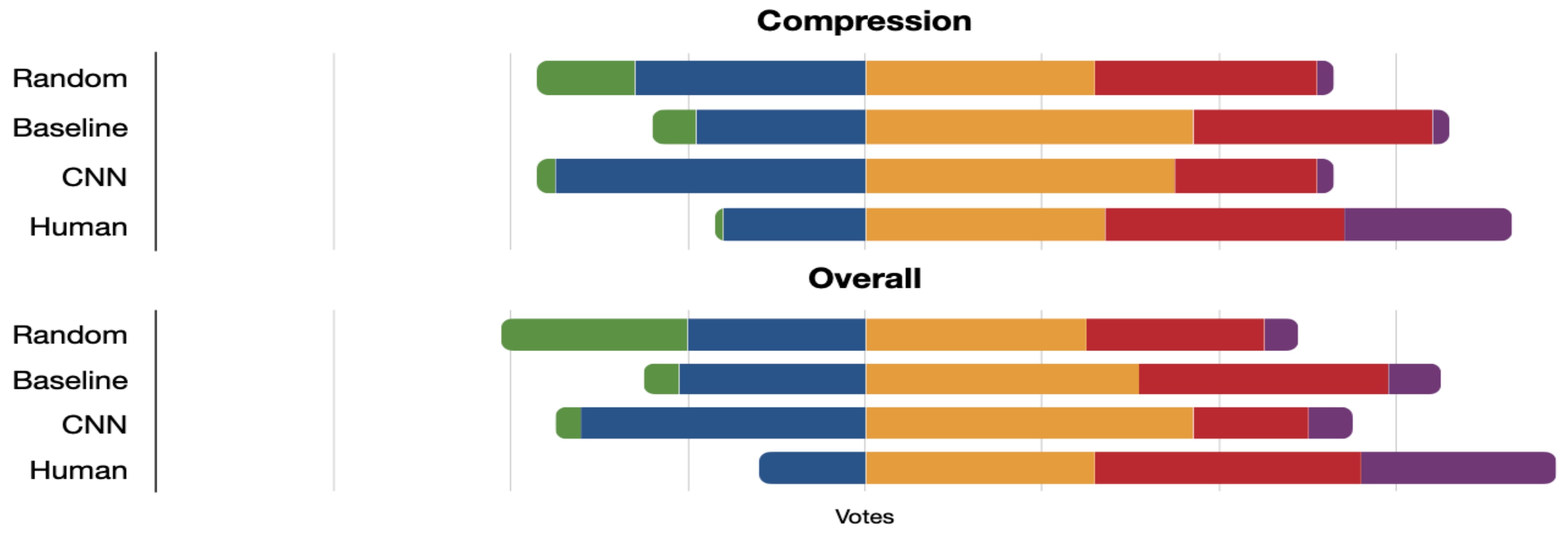
24 of the 30 participants claim to have experiences in mixing. 10 participants are audio professionals. Besides the mixes by the two systems, the test participants are presented with an human mix and a random mix.



None of the proposed systems outperforms the human mixes. There is no statistical significance between the two proposed systems.



The audio professionals rate the use compression much higher than the overall mixing quality in the above plot. The plot below is the rating by inexperienced participants. One explanation is that neither audio professionals nor inexperienced listeners can identify the usefulness of compression on vocals in wet mixes.



Some participants claim that they cannot judge the mixing quality when the source recording are are very good or very poor. Future analysis may help identify the true value of mixing in different scenarios. Furthermore, a follow-up investigation on the connection between the mixing settings and the perpetual mixing quality will be beneficial to the research field, the commercial mixing practices and audio education.

CONTACT

Kelian (Mike) Li
Music Informatics Group, Georgia Tech Center for Music Technology

kli421@gatech.edu
https://github.com/likelian/Lab_spring2022